Example-based Translation of Japanese Functional Expressions utilizing Semantic Equivalence Classes

Yusuke Abe† **Takafumi Suzuki**† **Bing Liang**† **Takehito Utsuro**†
Mikio Yamamoto† **Suguru Matsuyoshi**‡ **Yasuhide Kawada**††
†Graduate School of Systems and Information Engineering, University of Tsukuba,
Tsukuba, 305-8573, JAPAN
‡Department of Computer Science and Media Engineering,
Faculty of Engineering, University of Yamanashi,
4-3-11, Takeda, Kofu, Yamanashi, 400-8511, JAPAN
††Navix Co., Ltd., Tokyo, 141-0031, JAPAN

Abstract

This paper studies issues on machine translation of Japanese functional expressions into English. Unlike our previous works, in order to address the issue of resolving various ambiguities of a compound expression, this paper takes the approach of example-based machine translation. In this approach, a patent translation example database is developed given the phrase translation tables trained with parallel patent sentences as well as the training parallel patent sentences themselves. When identifying the most similar translation examples, we integrate semantic equivalence classes of Japanese functional expressions as well as more fine-grained similarity measure of translation examples. In the evaluation, we compare the translation accuracy of the proposed framework with that of Moses, and show that the proposed framework somehow outperforms Moses.

1 Introduction

The Japanese language has various types of functional expressions, which are very important for understanding their semantic contents. Those functional expressions are also problematic in further applications such as machine translation of Japanese sentences into English. This problem can be partially recognized by the fact that the Japanese language has a large number of variants of functional expressions, where their total number is recently counted as over 10,000 in Matsuyoshi et al. (2006). Based on those recent development in studies on lexicon for processing Japanese functional expressions (Matsuyoshi et al., 2006), this paper studies issues on machine translation of Japanese functional expressions into English.

In our previous works, Sakamoto et al. (2009) and Nagasaka et al. (2010) applied the "Sandglass" machine translation architecture (Yamamoto, 2002) to the task of translating Japanese functional expressions into English. In the "Sandglass" MT architecture, variant expressions in the source language are first paraphrased into representative expressions, and then, a small number of translation rules are applied to the representative expressions. In Sakamoto et al. (2009) and Nagasaka et al. (2010), we introduced the recently compiled large scale hierarchical lexicon of Japanese functional expressions (Matsuyoshi et al., 2006). We employed the semantic equivalence classes of the lexicon and examined each class whether it is monosemous or not. We realized this procedure by manually examining whether functional expressions within a class can be translated into a single canonical English expression. Then, we proposed how to extract rules for translating functional expressions in Japanese patent documents into English. Here, we used about 1.8M Japanese-English parallel sentences automatically extracted from Japanese-English patent families, which are distributed through the Patent Translation Task at the NTCIR-7 Workshop (Fujii et al., 2008). As a toolkit of a phrase-based SMT (Statistical Machine Translation) model, Moses (Koehn et al., 2007) was applied and Japanese-English translation pairs were obtained in the form of a phrase translation table. Finally, we extracted translation

Proceedings of the 4th Workshop on Patent Translation, pages 91-103,
MT Summit XIII, Xiamen, China, September 2011

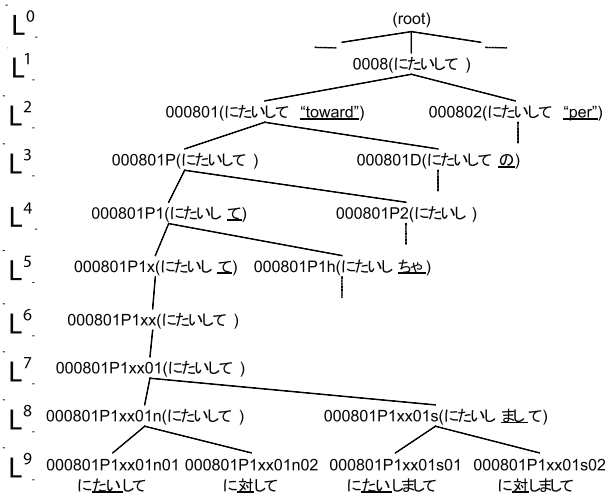

Figure 1: A Part of the Hierarchical Lexicon of Japanese Functional Expressions

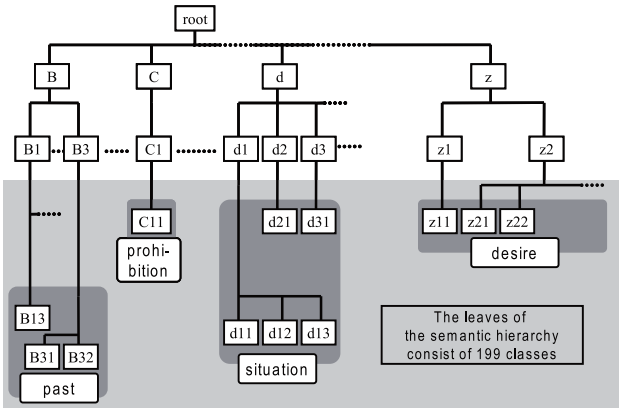

Figure 2: A Part of the Hierarchy of Semantic Equivalence Classes

pairs of Japanese functional expressions from the phrase translation table.

Unlike Sakamoto et al. (2009) and Nagasaka et al. (2010), in this paper, in order to address the issue of resolving various ambiguities of a compound expression in machine translation of Japanese functional expressions in patent documents, we take the approach of example-based machine translation (Sommers, 2003). In this approach, a patent translation example database is developed given the phrase translation tables trained with parallel patent sentences as well as the training parallel patent sentences themselves. When identifying the most similar translation examples, we integrate semantic equivalence classes of Japanese functional expressions as well as more fine-grained similarity measure of translation examples. In the evaluation of the proposed framework of example-based translation of Japanese functional expressions utilizing semantic equivalence classes, we compare the translation accuracy of the proposed framework with that of Moses (Koehn et al., 2007), and show that the proposed framework somehow outperforms Moses.

2 Hierarchical Lexicon of Japanese Functional Expressions

2.1 Morphological Hierarchy

In order to organize Japanese functional expressions with various surface forms, Matsuyoshi et al. (2006) proposed a methodology for compiling a lexicon of Japanese functional expressions with hierarchical organization[1]. Matsuyoshi et al. (2006) compiled the lexicon with 341 headwords and 16,801 surface forms. The hierarchy of the lexicon has nine abstraction levels and Figure 1 shows a part of the hierarchy[2]. In this hierarchy, the root node (in L^0) is a dummy node that governs all the entries in the lexicon. A node in L^1 is an entry (headword) in the lexicon; the most generalized form of a functional expression. A leaf node (in L^9) corresponds to a surface form (completely-instantiated form) of a functional expression. An intermediate node corresponds to a partially-abstracted (partially-instantiated) form of a functional expression. The second level L^2 distinguishes senses of Japanese functional expressions. This level enables distinction of more than one senses of one functional expression. For example, "にたいして" (ni-taishi-te) has two different senses. The first sense is "to"; e.g., "彼は私 にたいして 親切だ。" (He is kind to me). The second sense is "per"; e.g., "一人 にたいして5つ。" (five per one person). This level is introduced to distinguish such ambiguities. On the other hand, L^3 distinguishes grammatical functions, L^4 distinguishes alternations of function words, L^5 distinguishes phonetic variations, L^6 distinguishes optional focus particles, L^7 distinguishes conjugation

[1] http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

[2] In this lexicon, following Sag et al. (2002), each functional expression is regarded as a fixed expression, rather than a semi-fixed expression or a syntactically-flexible expression.

forms, L^8 distinguishes normal/polite forms, and L^9 distinguishes spelling variations.

2.2 Semantic Hierarchy

Along with the hierarchy of surface forms of functional expressions with nine abstraction levels, the lexicon compiled by Matsuyoshi et al. (2006) also has a hierarchy of semantic equivalence classes introduced from the viewpoint of paraphrasability. This semantic hierarchy has three abstraction levels, where 435 entries in L^2 (headwords with a unique sense) of the hierarchy of surface forms are organized into the top 45 semantic equivalence classes, the middle 128 classes, and the 199 bottom classes.

Figure 2 shows examples of the bottom 199 classes, where each of the leaf labels "B13", "B31", "B32", "C11", ..., "d11", "d12", "d13", ... represents a label of the bottom 199 classes. In Matsuyoshi and Sato (2008), the bottom 199 semantic equivalence classes of Japanese functional expressions are designed so that functional expressions within a class are paraphrasable in most contexts of Japanese texts.

3 Ambiguities of A Compound Expression

One of the key issues of this paper is whether each compound expression to be translated into English is monosemous or not. Unless each compound expression is monosemous, it is necessary to apply certain disambiguation techniques before translating it into English. Before we discuss how to consider ambiguities of compound expressions in the process of machine translation, this section overviews three types of ambiguities of compound expressions.

3.1 Ambiguity of Functional/Content Usages

The first type of ambiguity is for the case that one compound expression may have both a literal (i.e. compositional) *content word* usage and a non-literal (i.e. non-compositional) *functional* usage. This type of ambiguity often happens when the surface form of a functional expression can be decomposed into a sequence of at least one content word and one or more function words. In such a case, the surface form of the compound expression may have both a literal (i.e. compositional) *content word* usage where each of its constituents has its own literal usage, and a non-literal (i.e. non-compositional) *functional* usage where its constituents have no longer their literal usages.

For example, Table 2 shows two example sentences of a compound expression "もの (mono) の (no)", which consists of a formal noun "もの (mono)" and a post-positional particle "の (no)". In the sentence (2), the compound expression functions as an adversative conjunctive particle and has a non-compositional functional meaning "*although*". On the other hand, in the sentence (3), the expression simply corresponds to a literal concatenation of the usages of the constituents: the formal noun "もの (mono)" and the post-positional particle "の (no)", and has a literal meaning "*of*" for "の (no)", where the literal meaning of "もの (mono)" is omitted. Compared to Table 2, Table 1 shows an example of a functional expression without ambiguity of functional/content usages. In this case, the compound expression "こと (koto) が (ga) できる (dekiru)" consists of a formal noun "こと (koto)", a post-positional particle "が (ga)", and an auxiliary verb "できる (dekiru)". In almost all the occurrences in a newspaper corpus, the surface form of this compound expression functions as an auxiliary verb and has a non-compositional functional meaning "*can*".

3.2 Ambiguity of Functional Usages

The second type of ambiguity is for the case that the surface form of a functional expression has more than one *functional* usages. For example, Table 3 shows two example sentences of a compound expression "と (to) し (shi) て (te) も (mo)", which consists of a post-positional particle "と (to)", a conjunctive form "し (shi)" of a verb , a conjunctive particle "て (te)", and a post-positional particle "も (mo)". In the sentence (4), the compound expression belongs to a semantic equivalence class representing *adversative* sense, functions as a conjunctive particle and has a non-compositional functional meaning "". In the sentence (5), on the other hand, the compound expression belongs to a semantic equivalence class representing *topic*-mo-*perspective* sense, functions as a case-marking particle with a topic-marking particle *mo* (which means *"also"* in English) and has a non-compositional functional meaning "". Compared to Table 3, Table 1 shows an example of a functional expression without ambigu-

Ambiguities of A Compound Expression (a):
Table 1: *w/o* ambiguity of functional usages NOR *w/o* ambiguity of functional/content usages NOR *w/o* ambiguity of translation into English

	Expression	Example sentence (English translation)	Usage
(1)	ことができる (koto-ga-dekiru)	また設計上の必要に応じて第１および第２の部材２２，２３の寸法形状を変えれば、設計上の適正値を実現する ことができる 。 (Further by changing the size and configuration of the first and second members 22 and 23 in accordance with any requirements in design, *it is possible* to realize proper design values.)	functional, semantic class = *possible* (ことができる (koto-ga-dekiru) = *it is possible*)

Table 2: Ambiguities of A Compound Expression (b): *with* ambiguity of functional/content usages

	Expression	Example sentence (English translation)	Usage
(2)	ものの (mono-no)	乾燥に供した加熱空気は蒸発した水蒸気を含み，多くの熱エネルギーを持っている ものの ，回収して循環利用するには限界があり，多くの場合廃棄されている． (*Although* the heated air provided for drying contains the evaporated water vapor and has high heat energy, it is limited in terms of the recovery and circulation for re-use and hence is discarded in many cases.)	functional, semantic class = *adversative* (～ ものの (mono-no) = *although* ～
(3)	ものの (mono-no)	ここで，ブロックが存在しない場合は，探索対象段の位置を，保持されたアベイラブルエリアで最後の ものの 左上隅点とし（ステップ１１０６），その後，後述する図１２に示される処理を実行する。 (If not, the available area generating unit 108 sets the position of the column to be searched to the top left corner point *of* the last available area element in the held available area (step 1106) and then executes the processing shown in Fig 12 to be described later in more detail.)	content (～ ものの) (mono-no))) = *of*)

ity of functional usages. In this case, the functional expression “こと (koto) が (ga) できる (dekiru)” has only one non-compositional functional meaning “*can*”.

This type of ambiguity includes issues on typical polysemies and homographs, where the issues on sense disambiguation of content words have been well studied in NLP community (e.g, in SENSEVAL tasks (Kilgarriff and Palmer, 2000; Kurohashi and Uchimoto, 2003)). However, in the areas of semantic analysis of Japanese sentences as well as machine translation of Japanese sentences, the issue of sense disambiguation of functional expressions has not been paid much attention so far, and any standard tool for sense disambiguation of Japanese functional expressions have not been publicly available.

3.3 Ambiguity of Translation into English

The third type of ambiguity is for the case that the surface form of a functional expression has a single usage, while its translation into English are different depending on its usage within a sentence. For example, Table 4 shows three example sentences of a compound expression “に (ni) よる (yoru)”, which consists of a post-positional particle “に (ni)” and a base form “よる” of a verb, and belongs to a semantic equivalence class representing *cause* sense. Its translation into English is “*by*” in the sentence (6), “*according to*” in the sentence (7), and “*due to*” in the sentence (8). In (6), the expression represents a

Table 3: Ambiguities of A Compound Expression (c): *with* ambiguity of functional usages

	Expression	Example sentence (English translation)	Usage
(4)	としても (to-shi-te-mo)	このため、誤って装置に物等を落下した としても, その衝撃は反射ミラー８ｆに伝わり難くなっている. (With this arrangement in place *even when* the something is dropped on the apparatus by mistake, its impact is unlikely to be transmitted to the reflection mirror 8f.)	functional, semantic class = *adversative* (としても (to-shi-te-mo) = *even when*)
(5)	としても (to-shi-te-mo)	さらに，ブレード４５は接触ローラ３７の外周面３７ａの汚れを除去するクリーニング手段 としても 作用する. (Furthermore, the blade 45 functions *as* a cleaning means for removing dirts on the circumference 37a of the contact roller 37.)	functional, semantic class = *topic*-mo-*perspective* (としても (to-shi-te-mo) = *as*)

meaning like "something is caused *by* something". In (7), it represents a meaning like "some function is provided *according to* the embodiment of some invention". Finally, in (8), it represents a meaning like "some effect is obtained *due to* some action".

4 Developing Patent Translation Example Databases

4.1 Japanese-English Parallel Patent Documents

In the Japanese-English patent translation task of the NTCIR-7 workshop (Fujii et al., 2008), parallel patent documents and sentences were provided by the organizer. Those parallel patent documents are collected from the 10 years of unexamined Japanese patent applications published by the Japanese Patent Office (JPO) and the 10 years patent grant data published by the U.S. Patent & Trademark Office (USPTO) in 1993-2000. The numbers of documents are approximately 3,500,000 for Japanese and 1,300,000 for English. Because the USPTO documents consist of only patent that have been granted, the number of these documents is smaller than that of the JPO documents.

From these document sets, patent families are automatically extracted and the fields of "Background of the Invention" and "Detailed Description of the Preferred Embodiments" are selected. This is because the text of those fields is usually translated on a sentence-by-sentence basis. Then, the method of Utiyama and Isahara (2007) is applied to the text of those fields, and Japanese and English sentences are aligned. The number of Japanese and English parallel sentences is about 1.8M in total.

4.2 Phrase Translation Table of an SMT Model

As a toolkit of a phrase-based statistical machine translation model, we use Moses (Koehn et al., 2007) and apply it to the whole 1.8M parallel patent sentences. In Moses, first, word alignment of parallel sentences are obtained by GIZA++ (Och and Ney, 2003) in both translation directions and then the two alignments are symmetrised. Next, any phrase pair that is consistent with word alignment is collected into the phrase translation table and a phrase translation probability is assigned to each pair (Koehn et al., 2003). We finally obtain 76M translation pairs with 33M unique Japanese phrases, i.e., 2.29 English translations per Japanese phrase on average, with Japanese to English phrase translation probabilities $P(p_E \mid p_J)$ of translating a Japanese phrase p_J into an English phrase p_E. For each Japanese phrase, those multiple translation candidates in the phrase translation table are ranked in descending order of Japanese to English phrase translation probabilities.

4.3 The Procedure of Developing A Translation Example Database per Semantic Equivalence Class

Given the phrase translation tables trained with parallel patent sentences as well as the training parallel patent sentences themselves, a patent translation example database is developed according to the proce-

Table 4: Ambiguities of A Compound Expression (d): *with* ambiguity of translation into English

	Expression	Example sentence (English translation)	Usage
(6)	による (ni-yoru)	原稿台１１側からの光のミラー１４ による 反射光路上には結像レンズ１６とプラテン２０がこの順に配置されている。 (A lens 16 and a platen 20 are arranged (in the named order) in a path to pass the light reflected *by* the mirror 14 from the original table 11.)	functional, semantic class = *cause* (による (ni-yoru) = *by*)
(7)	による (ni-yoru)	本発明 による 可変差動制限装置２の制御は、以下の(1)，(2)，(3) の３種の制御の組合せから構成される。 (The control of the variable differential motion limiting device 2 *according to* the embodiment of the present invention comprises a combination of the following three controls (1), (2), and (3).)	functional, semantic class = *cause* (による (ni-yoru) = *according to*)
(8)	による (ni-yoru)	つまり、放電開始 による 電圧の低下が、極間異常状態と判定されてしまうことがある。 (Namely, the voltage reduction *due to* the discharge start may sometimes be judged as an abnormal machining gap status.)	functional, semantic class = *cause* (による (ni-yoru) = *due to*)

dure illustrated in Figure 3.

First of all, in our framework of example-based translation of Japanese functional expressions, we design the translation example database as having one example database for each of the 199 semantic equivalence classes. Furthermore, following the notion of "Sandglass" machine translation architecture, we restrict translation examples to be included in the example database as only the representative ones. To realize this, we have the lower bound of the phrase translation probability as 0.05, that of the phrase translation frequency as 10, and that of the frequency of a Japanese compound expression as 20.

Then, for compound expressions of each semantic equivalence class, we collect translation example patent sentences from those used for training phrase translation tables. Here, we only keep translation examples with a phrase translation pair which satisfies the lower bounds of probability / frequencies above. Finally, in the preliminary evaluation in section 6, we simply select five translation example sentences for each compound expression and collect them into the translation example databases.

5 Example-based Translation of Japanese Functional Expressions utilizing Semantic Equivalence Classes

In our framework of example-based translation of Japanese functional expressions, as is the case of standard example-based machine translation frameworks (Sommers, 2003), given a sentence to be translated, we search the translation example database for the most similar translation example sentence pair. One exception is, however, that we have one example database for each semantic equivalence class, and, given a compound expression belonging to a semantic equivalence class and a Japanese patent sentence including the expression, we search for the most similar translation example only within the example database of the corresponding semantic equivalence class, but not the whole example databases.

In our framework, an example e of translation sentences collected in the translation example databases is represented as shown in Table 5. Given a Japanese compound expression, its phrase translation pair is obtained from the phrase translation table. For example, in the case of the example in Table 5, a compound expression "と (to) し (shi) て

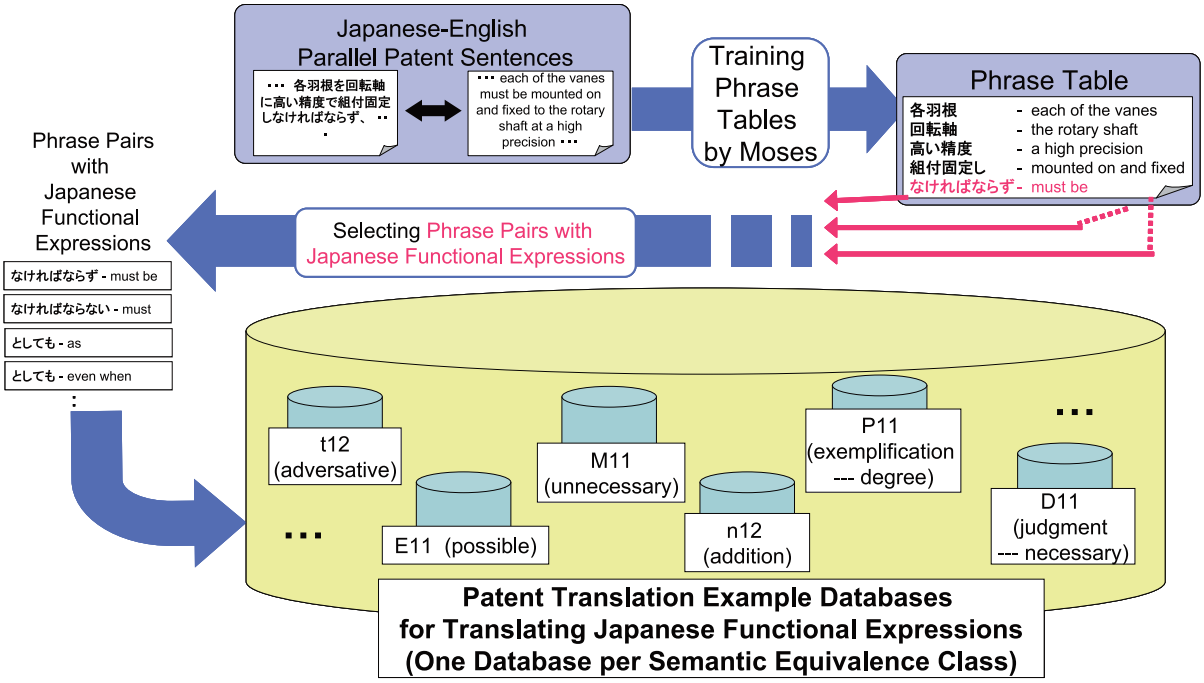

Figure 3: Procedure of Developing A Translation Example Database per Semantic Equivalence Class

(te) も (mo)" is given and a translation example sentence pair with a phrase alignment "と (to) し (shi) て (te) も (mo)" and "*even if*" found in the phrase translation table are shown[3].

When searching for the most similar translation example throughout the example database, a certain similarity measure is employed. In this paper, we first represent the Japanese sentence of e as a tuple of $\langle m_{pre}, M_c, m_{suf} \rangle$, where m_{pre} and m_{suf} denote the morpheme preceding the compound expression and the one subsequent to it, while M_c denotes the sequence of morphemes constituting the compound expression. Then, we define the similarity measure $Sim(e_1, e_2)$ of two examples e_1 and e_2 as below:

$$\begin{aligned} Sim(e_1, e_2) = & \\ & Sim_{pre}(m_{pre}(e_1), m_{pre}(e_2)) \\ & + Sim_c(M_c(e_1), M_c(e_2)) \\ & + Sim_{suf}(m_{suf}(e_1), m_{suf}(e_2)) \end{aligned}$$

where Sim_{pre} and Sim_{suf} denote the similarity measure of preceding as well as subsequent morphemes. Sim_c denotes the similarity measure of the morpheme sequence constituting the compound expression. More precisely, those component similarity measures count the number of identical parts-of-speech tags and conjugation forms between $m_{pre}(e_1)$ and $m_{pre}(e_2)$, $m_{suf}(e_1)$ and $m_{suf}(e_2)$, and corresponding constituents of $M_c(e_1)$ and $M_c(e_2)$[4] [5].

[3]Note that, in each translation example database of a semantic equivalence class, we do not exclude any type of ambiguities shown in Table 2 (ambiguity of functional/content usages), Table 3 (ambiguity of functional usages), and Table 4 (ambiguity of translation into English). For each surface form belonging to a semantic equivalence class C, together with the translation examples with the functional usage of the class C, we may include translation examples with content usages as well as functional usages of other semantic equivalence classes in the translation example database of the class C. This means that we do not include any result of semantic disambiguation into the translation example databases in advance, but that all of the semantic disambiguation processes are to be done only through the example-based translation process.

[4]In Sim_c, we only align prefix sequences of constituent morpheme sequences of $M_c(e_1)$ and $M_c(e_2)$, and count the number of identical parts-of-speech tags and conjugation forms, where we ignore the suffix sequence of the longer sequence between $M_c(e_1)$ and $M_c(e_2)$.

[5]Sometimes, it can happen that there exist more than one most similar translation examples in the example database, and English translations of those examples are semantically different. In such cases, we incorporate fine-grained similarity mea-

Table 5: An Example e of the Translation Sentences in the Translation Example Database

Japanese sentence	このため、どのような体格の乗員が、どのような着座姿勢で、どのような衣服を着用していた としても 、その場合に応じた緩み量に見合った量の引き込みを行うことができる。
Japanese compound expression in context	preceding morpheme $m_{pre}(e)$: た (ta) / ⟨auxiliary verb, base form⟩ constituent morpheme sequence $M_c(e)$: と (to)/⟨post-positional particle⟩ - し (shi)/⟨verb, conjunctive form⟩ -て (te)/⟨conjunctive particle⟩ - も (mo)/⟨topic-marking particle⟩ subsequent morpheme $m_{suf}(e)$: 、/⟨comma⟩
English translation	As a result, even if a passenger with any build takes any seating posture and has any clothes on, retraction of the webbing 1 can be carried out by an amount corresponding to the slack amount.
English phrase	even if

Table 6: Evaluation Results

Semantic Equivalence Class	# of expressions in the database	# of examples in the database	# of sentences for evaluation	translation accuracy (%)	
				Moses	Proposed
c11 (cause)	5	56	15	87	**100**
M11 (unnecessary)	5	66	15	47	**93**
m12 (confinement)	2	22	15	100	73
n12 (addition)	4	31	15	93	80
s11 (reason — situational)	2	15	15	87	80
t12 (adversative)	5	37	15	40	**67**
P11 (exemplification — degree)	3	12	15	53	47
total	26	239	105	72	**77**

6 Evaluation

In the evaluation of the proposed framework of example-based translation of Japanese functional expressions utilizing semantic equivalence classes, we focus on semantic equivalence classes which have relatively high ambiguities of compound expressions presented in section 3. Then, we compare the translation accuracy of the proposed framework with that of Moses (Koehn et al., 2007), and show that the proposed framework somehow outperforms Moses.

sure of functional expressions based on the morphological hierarchy presented in section 2.1 and select the most similar one.

6.1 The Procedure

Out of the total 199 semantic equivalence classes, for 178 classes, at least one compound expression is included in the training 1.8M parallel sentences. Then, in Nagasaka et al. (2010), we examined 53 classes out of the 178, whether or not each class has relatively high ambiguities of compound expressions presented in section 3. Finally, we select seven classes listed in Table 6. Then, according to the procedure we presented in section 4.3, for each of the seven classes, we collect compound expressions and translation example sentence pairs which satisfy the lower bounds of probability / frequencies. Then, we develop a translation example database for each of the seven classes. Table 6 lists the numbers of compound expressions as well as translation exam-

Table 7: Analysis on Coverage of the Example Database and Ambiguity of A Compound Expression

			rate (%)	translation accuracy (%)	
				Moses	Proposed
Coverage of the Example Database	expression *is covered*	compound expression is *unambiguous*	52	75	78
		compound expression is *ambiguous*	16	94	82
	expression *is not covered*		32	58	**72**

ple sentence pairs for each class.

Next, for each compound expression of the seven classes, from the Japanese-English parallel patent sentence pairs of the years during 2001-2007 provided at the patent translation task of the NTCIR-8 workshop (Fujii et al., 2010), we collect parallel sentences for evaluation (the number of parallel sentences is as shown in Table 6).

6.2 Evaluation Results

Table 6 shows the results of comparing translation accuracies between the proposed framework and Moses for each of the seven classes. Overall, the proposed framework outperforms Moses by about 5%. For three out of the seven classes, the proposed framework outperforms Moses, while for the remaining four classes, Moses outperformed the proposed framework.

Table 7 analyzes the coverage of the translation example databases as well as the ambiguity of each compound expression for evaluation.

First, in 32% of the evaluation sentences, example databases do not include translation example sentence pairs which have exactly the same compound expression that is included in the evaluation sentence. Translation accuracy of the proposed framework is relatively higher than that of Moses mostly in this case. In such cases, those compound expressions have relatively low frequencies, and thus their translation in the phrase table tend not to be reliable.

Table 8 shows an example of this case and compares it with translation by Moses. In this example, the compound expression "に (ni) しろ (shiro)", which consists of a post-positional particle "に (ni)" and an imperative form "しろ (shiro)" of a verb, belongs to a semantic equivalence class representing *adversative* sense. This expression has relatively low frequency in the training parallel patent sentences, and the example database of this class does not include translation examples including this expression. Then, by our example-based translation framework, a translation example having a more frequent compound expression "と (to) し (shi) て (te) も (mo)", which belongs to the same class, is selected as the most similar translation example. Finally, from the most similar translation example, "*even if*", a typical English translation of *adversative* sense, is found as the translation by the proposed framework. On the other hand, as can be seen from Table 8, translation by Moses is damaged in generating an English phrase representing *adversative* sense.

Second, for the remaining 68% evaluation sentences, the example databases include translation example sentence pairs which have exactly the same compound expression that is included in the evaluation sentence. Out of those cases, we next focus on 16% of them where compound expressions in each example database are ambiguous in terms of the ambiguities presented in section 3. In those cases, the proposed framework and Moses perform evenly well in resolving ambiguities of compound expressions.

Table 9 shows an example of this case and again compares it with translation by Moses. In this example, the compound expression included is "と (to) し (shi) て (te) も (mo)", which has the ambiguity of functional usages as shown in Table 3. By our example-based translation framework, out of the two functional usages listed in Table 3, a translation example having *adversative* sense is selected as the most similar translation example. On the other hand, as can be seen from Table 9, translation by Moses is again damaged in generating an English phrase representing *adversative* sense.

Compared with those cases where the proposed method outperforms Moses, the opposite cases where Moses outperforms the proposed method can be roughly categorized into the followings: (i) the

Table 8: An Example where the Proposed Method Outperforms Moses
(for the Semantic Equivalence Class: t12 (adversative))
(a) Identical expression is not included in the example database,
where an example of a more frequent expression in the semantic equivalence class is selected.

Japanese sentence for evaluation (reference translation)	それは膜中において一様に分布していない にしろ 、平均的な濃度とすれば、１×１０１９／ｃｍ３を越える濃度で残存している。 (Even if the metallic element is not distributed uniformly within the film, it remains at an average concentration that exceeds 1.times.10.sup.19 / cm.sup.3.)

The most similar translation example in the database	このため、どのような体格の乗員が、どのような着座姿勢で、どのような衣服を着用していた としても 、その場合に応じた緩み量に見合った量の引き込みを行うことができる。 As a result, even if a passenger with any build takes any seating posture and has any clothes on, retraction of the webbing 1 can be carried out by an amount corresponding to the slack amount.
Similarity calculation	Sim_{pre}(ない (nai)/⟨auxiliary verb, base form⟩, た (ta)/⟨auxiliary verb, base form⟩) = 3/3 = 1 Sim_c(に (ni)/⟨post-positional particle⟩ -しろ (shiro)/⟨verb, imperative form⟩, と (to)/⟨post-positional particle⟩ -し (shi)/⟨verb, conjunctive form⟩ -て (te)/⟨conjunctive particle⟩ -も (mo)/⟨topic-marking particle⟩) = (3/3 + 2/3) / 2 = 0.83 Sim_{suf} (、/⟨comma⟩, 、/⟨comma⟩) = 3/3 = 1 $Sim = Sim_{pre} + Sim_c + Sim_{suf} = 2.83$

Translation by Moses	In the film not uniformly distributed in this case, if the average density and a concentration exceeding 1.times.10.sup.19 / cm.sup.3 remains.

proposed method selects a translation example with a translation probability lower than the maximum translation probability and is judged as *incorrect*, while Moses selects an English translation with the maximum translation probability and is judged as *correct*. (ii) Moses selects a phrase translation table entry with a Japanese compound expression that is longer than the one the proposed method selects, where only the translation by Moses is judged as *correct*. (iii) Moses skips to translate the Japanese compound expression into English, where only the translation by Moses is judged as *correct*.

Table 10 shows an example of the category (ii) above. In this example, the compound expression included is “ばかり (bakari)”. By our example-based translation framework, a translation example with translation into English as “*about*” is selected as the most similar translation example. On the other hand, Moses selects a phrase translation table entry with a compound expression “ばかり (bakari) で (de) なく (naku)” that is longer than the one the proposed method selects, and with translation into English as “*but also*”. The Japanese sentence is a typical example of a functional usage of the compound expression “ばかり (bakari) で (de) なく (naku)”, and only the translation by Moses is judged as *correct*.

7 Related Works

Ambiguities of functional/content usages has been well studied in Tsuchiya et al. (2005), Tsuchiya et al. (2006), and (Shudo et al., 2004). Tsuchiya et al. (2005) reported that, out of about 180 compound expressions which are frequently observed in the newspaper text, one third (about 60 expressions) have this type of ambiguity. Next, Tsuchiya et al. (2006) formalized the task of identifying Japanese compound functional expressions in a text as a machine learning based chunking problem. The proposed technique performed reasonably well, while its major drawback is in its scale. So far, the proposed technique has not yet been applied to the whole list of over 10,000 Japanese functional ex-

Table 9: An Example where the Proposed Method Outperforms Moses
(for the Semantic Equivalence Class: t12 (adversative))
(b) Disambiguation of Functional Usages

Japanese sentence for evaluation (reference translation)	しかも、演奏データのチャンネル配置を変更した[としても]、結果的に発音される演奏音については何も変更されない（どのチャンネルを使用して発音しようとも、指定された音色等で発音がなされればよい）ので、透し情報を埋め込んだことによる再生演奏への悪影響は全く生じず、極めて有利である。 (In addition, even though the assignment of the performance data sets to the channels is changed in the above mentioned manner, no change is made to performance tones to be eventually sounded (it is only necessary that the performance tones be sounded with designated tone colors etc. no matter which channels are used for the sounding); thus, embedding the electronic watermark information will, in no way, adversely influence the reproductive performance of the tones and the inventive technique will prove extremely advantageous.)
The most similar translation example in the database	このため、どのような体格の乗員が、どのような着座姿勢で、どのような衣服を着用していた[としても]、その場合に応じた緩み量に見合った量の引き込みを行うことができる。 As a result, [even if] a passenger with any build takes any seating posture and has any clothes on, retraction of the webbing 1 can be carried out by an amount corresponding to the slack amount.
Similarity calculation	Sim_{pre}(た (ta)/⟨auxiliary verb, base form⟩, た (ta)/⟨auxiliary verb, base form⟩) = 3/3 = 1 Sim_c(と (to)/⟨post-positional particle⟩ - し (shi)/⟨verb, conjunctive form⟩ -て (te)/⟨conjunctive particle⟩ - も (mo)/⟨topic-marking particle⟩, と (to)/⟨post-positional particle⟩ - し (shi)/⟨verb, conjunctive form⟩ -て (te)/⟨conjunctive particle⟩ - も (mo)/⟨topic-marking particle⟩) = (3/3 + 3/3 + 3/3 + 3/3) / 4 = 1 Sim_{suf} (、/⟨comma⟩, 、/⟨comma⟩) = 3/3 = 1 $Sim = Sim_{pre} + Sim_c + Sim_{suf} = 3.0$
Translation by Moses	In addition, a channel of the performance data is changed, as a result of play tones are sounded, nothing is not changed (which channel is used both for the designated tone color, etc. to be sounded tone generation is performed by the information may be affected to playing is embedded is very advantageous does not occur at all.

pressions. (Shudo et al., 2004) also studied applying manually created rules to the task of resolving functional/content ambiguities, where their approach has limitation in that it requires human cost to create manually and to maintain those rules.

Utsuro et al. (2007) and (Nivre and Nilsson, 2004) studied syntactic analysis of functional expressions in sentences. Utsuro et al. (2007) studied how to incorporate the process of analyzing compound non-compositional functional expressions into the framework of Japanese statistical dependency parsing. (Nivre and Nilsson, 2004) also reported improvement of Swedish parsing when multi word units are manually annotated.

8 Concluding Remarks

This paper studied issues on machine translation of Japanese functional expressions into English. Unlike our previous works, in order to address the issue of resolving various ambiguities of a compound expression, this paper took the approach of example-based machine translation. When identifying the most similar translation examples, we integrated semantic equivalence classes of Japanese functional expressions as well as more fine-grained similarity measure of translation examples. In the evaluation,

Table 10: An Example where Moses Outperforms the Proposed Method
(for the Semantic Equivalence Class: P11 (exemplification — degree))

Japanese sentence for evaluation (reference translation)	上記従来の薬液持続注入器は、負圧シリンダの気密を維持するためにシリンダ及びシリンジ内の２個所に摺動抵抗の大きな個所を必要とし、精密製造するためには困難が伴った ばかり でなく、使用上も薬液注入に大きな力が必要となり操作性が悪いという欠点があった。 (The conventional continuous liquid infusion device as described above requires two positions having large sliding resistance in the cylinder and the syringe in order to maintain air tightness of the negative pressure cylinder, which causes difficulty in precise manufacturing as well as a disadvantage of poor operability in usage because large force is needed for infusing liquid.)

The most similar translation example in the database	この場合、本来は予測に用いないフレームから予測されることになるので、画質の劣化が予想されるが、一般的な画像では順方向モードの割合が１５％ ぐらい なので、大きな問題とはならない。 In the ordinary image, however, since the percentage of the forward direction mode is about 15%, a big problem will not arise.
Similarity calculation	Sim_{pre}(た (ta)/⟨auxiliary verb, base form⟩, % (%)/⟨noun-suffix⟩) = 0/3 = 0 Sim_c(ばかり (bakari)/⟨topic-marking particle⟩, ぐらい (gurai)/⟨topic-marking particle⟩) = 3/3 = 1 Sim_{suf} (で (de)/⟨auxiliary verb, conjunctive form⟩, な (na)/⟨auxiliary verb, adnominal form⟩) = 2/3 = 0.66 $Sim = Sim_{pre} + Sim_c + Sim_{suf} = 1.66$

Translation by Moses	The above conventional continuous dosing of negative pressure cylinder, in order to maintain the airtightness of the cylinder and the syringe 24 in a large sliding resistance at two positions on the precision is required for manufacturing, but also it is used with a large force is required on the chemical implantation of poor operability.

we compared the translation accuracy of the proposed framework with that of Moses, and showed that the proposed framework somehow outperforms Moses (Koehn et al., 2007). Future works include scaling up the translation example databases to all of the 199 semantic equivalence classes. It is also interesting to integrate the results of the proposed framework with those by Moses through a machine learning framework.

References

A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proc. 7th NTCIR Workshop Meeting*, pages 389–400.

A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, H. Echizen-ya, T. Ehara, and S. Shimohata. 2010. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proc. 8th NTCIR Workshop Meeting*, pages 371–376.

A. Kilgarriff and M. Palmer. 2000. Introduction to the special issue on SENSEVAL. *Computers and Humanities*, 34:1–13.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pages 127–133.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pages 177–180.

S. Kurohashi and K. Uchimoto. 2003. SENSEVAL-2 Japanese Translation Task. *Journal of Natural Language Processing*, 10(3):25–37.

S. Matsuyoshi and S. Sato. 2008. Automatic paraphrasing of Japanese functional expressions using a hierarchically organized dictionary. In *Proc. 3rd IJCNLP*, pages 691–696.

S. Matsuyoshi, S. Sato, and T. Utsuro. 2006. Compilation of a dictionary of Japanese functional expressions with hierarchical organization. In *Proc. ICCPOL*, LNAI: Vol. 4285, pages 395–402. Springer.

T. Nagasaka, R. Shimanouchi, A. Sakamoto, T. Suzuki, Y. Morishita, T. Utsuro, and S. Matsuyoshi. 2010. Utilizing semantic equivalence classes of Japanese functional expressions in translation rule acquisition from parallel patent sentences. In *Proc. 7th LREC*, pages 1778–1785.

J. Nivre and J. Nilsson. 2004. Multiword units in syntactic parsing. In *Proc. LREC Workshop, Methodologies and Evaluation of Multiword Units in Real-World Applications*, pages 39–46.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. 3rd CICLING*, pages 1–15.

A. Sakamoto, T. Nagasaka, T. Utsuro, and S. Matsuyoshi. 2009. Identifying and utilizing the class of monosemous Japanese functional expressions in machine translation. In *Proc. 23rd PACLIC*, pages 803–810.

K. Shudo, T. Tanabe, M. Takahashi, and K. Yoshimura. 2004. MWEs as non-propositional content indicators. In *Proc. 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 32–39.

H. Sommers. 2003. An overview of EBMT. In M. Carl and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 3–57. Kluwer Academic.

M. Tsuchiya, T. Utsuro, S. Matsuyoshi, S. Sato, and S. Nakagawa. 2005. A corpus for classifying usages of Japanese compound functional expressions. In *Proc. PACLING*, pages 345–350.

M. Tsuchiya, T. Shime, T. Takagi, T. Utsuro, K. Uchimoto, S. Matsuyoshi, S. Sato, and S. Nakagawa. 2006. Chunking Japanese compound functional expressions by machine learning. In *Proc. Workshop on Multi-Word-Expressions in a Multilingual Context*, pages 25–32.

M. Utiyama and H. Isahara. 2007. A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pages 475–482.

T. Utsuro, T. Shime, M. Tsuchiya, S. Matsuyoshi, and S. Sato. 2007. Learning dependency relations of Japanese compound functional expressions. In *Proc. Workshop on A Broader Perspective on Multiword Expressions*, pages 65–72.

K. Yamamoto. 2002. Machine translation by interaction between paraphraser. In *Proc. 19th COLING*, pages 1107–1113.